

# Simpson's Paradox and Collapsibility

P. Vellaisamy

*Department of Mathematics, Indian Institute of Technology Bombay,  
Powai, Mumbai-400 076, India*

**Abstract.** Simpson's paradox and collapsibility are two closely related concepts in the context of data analysis. While the knowledge about the occurrence of Simpson's paradox helps a statistician to draw correct and meaningful conclusions, the concept of collapsibility deals with dimension-reduction aspects, when Simpson's paradox does not occur. We discuss in this paper in some detail the nature and the genesis of Simpson's paradox with respect to well-known examples and also various concepts of collapsibility. The main aim is to bring out the close connections between these two phenomena, especially with regard to the analysis of contingency tables, regression models and a certain measure of association or a dependence function. There is a vast literature on these topics and so we focus only on certain aspects, recent developments and some important results in the above-mentioned areas.

**Key words:** *Collapsibility, contingency table, regression models, Simpson's paradox.*

## 1 Introduction

It is well known that statistics or more precisely statistical techniques play an important role in addressing some of the problems of a society, an industry and a country. But, drawing intelligent and correct decisions from the real-life data is not straightforward. Indeed, we read/hear different or paradoxical conclusions in several contexts or situations. So, a layman gets confused and probably believes in the famous quote "Lies, Damned Lies and Statistics," in that order. In this paper, we will address one of the well-known paradoxes due to Simpson. Simpson (1951) discussed how a simple fact about fractions can lead to a contradictory conclusion in a wide variety of situations. Though some statisticians (Yule (1903)) were aware of these issues in the beginning of 20th century, it is Simpson who popularized this paradox, earning his name, through analysis of real-life data that arise in several practical applications.

Simpson's paradox occurs when an observed association between two random variables, say  $X$  and  $Y$ , gets reversed after considering the third variable  $W$ , called a covariate or a background variable. The situation of having two contradictory conclusions makes this phenomenon paradoxical.

It is one of the most discussed and studied paradoxes in the statistics literature. The knowledge and the awareness of its occurrence is of importance for the statistical analysis of the data. It arises naturally in several areas which include the analysis of contingency tables, regression models, measures of association, survival analysis, etc. In this paper, we will discuss some examples of Simpson's paradox through some real-life data and then discuss some recent results in the areas mentioned above. A closely related concept, namely collapsibility, is applied whenever the Simpson's paradox does not occur. We present a survey of some recent results, its relation to Simpson's paradox and discuss possible directions for future work.

## 2 Simpson's Paradox

### 2.1 Real-life situations

We start with two examples discussed in the literature.

**Example 1** (Graduate Admission Data in UC Berkeley). Let  $A \in \{Y, N\}$  denote the admission,  $X \in \{M, F\}$  denote the sex and  $D \in \{H, G\}$  denote the department (H=History, G = Geography). The following data roughly represents the graduate admission in the two departments  $H$  and  $G$  (Bickel *et. al* (1975), discrimination suit against UCB ).

		D	
A	X	H	G
Y	M	1	6
	F	2	4
N	M	4	2
	F	6	1

We observe

$$P(Y|M, H) = \frac{1}{5} < \frac{2}{8} = P(Y|F, H); \quad (1)$$

$$P(Y|M, G) = \frac{6}{8} < \frac{4}{5} = P(Y|F, G). \quad (2)$$

That is, departmentwise women applicants are favored and hence there is no bias against them. But, considering the marginal table for  $A$  and  $X$ ,

		X	
		M	F
A	Y	7	6
	N	6	7

we get

$$P(Y|M) = \frac{7}{13} > \frac{6}{13} = P(Y|F), \quad (3)$$

showing that overall men do better than women. The reversal of the inequality in (3), contrary to the ones in (1)-(2), is called Simpson's paradox.

Why does this paradox occur? To answer this question, we need to look for additional information the data contains. First, look at the marginal table between  $A$  and  $D$ :

		D	
		H	G
A	Y	3	10
	N	10	3

We have

$$P(Y|H) = \frac{3}{13} < \frac{10}{13} = P(Y|G),$$

showing that getting admission in history is tougher than in geography.

Next, look at the marginal table between  $X$  and  $D$ :

		D	
		H	G
X	M	5	8
	F	8	5

It is clear that

$$P(H|M) = \frac{5}{13} < \frac{8}{13} = P(H|F),$$

because more women applied to history department than in geography. Since the admission in history is tougher than in geography, the reversal in (3) occurs.

A probabilistic justification due to Blyth (1972) is the following. Note that

$$\begin{aligned} P(Y|M) &= P(Y|M, H)P(H|M) + P(Y|M, G)P(G|M) = \frac{1}{5} \cdot \frac{5}{13} + \frac{6}{8} \cdot \frac{8}{13} = \frac{7}{13} \\ &= E_{D|M}P(Y|M, D) \\ P(Y|F) &= P(Y|F, H)P(H|F) + P(Y|F, G)P(G|F) = \frac{2}{8} \cdot \frac{8}{13} + \frac{4}{5} \cdot \frac{5}{13} = \frac{6}{13} \\ &= E_{D|F}P(Y|F, D) \end{aligned}$$

Because of  $\frac{5}{13} = P(H|M) < P(H|F) = \frac{8}{13}$ , the reversal  $P(Y|M) > P(Y|F)$  occurs. Observe that the conditional distribution  $\mathcal{L}(D|X)$  also plays a key role in Simpson's paradox.

**Example 2** Consider the following data (Agresti, (1990)) concerning death penalty (D), race of the accused (A) and race of the victim (V). Also, let  $W$  and  $B$  denote the white and the black, respectively.

		D	
		Y	N
W	W	19	132
	B	0	9
B	W	11	52
	B	6	97

First look at the marginal table corresponding to  $A$  and  $D$ . Let  $A_W$  or  $A_B$  denote that accused is a  $W$  or  $B$  (similar meaning for  $V_W$  and  $V_B$ ).

		D	
		Y	N
A	W	19	141
	B	17	149

From the above table,

$$P(Y|A_W) = 19/160 = 0.12 > 17/166 = 0.10 = P(Y|A_B),$$

showing that the white accused are more likely to get death penalty. However, if we consider the victim's race also, we have from the first table that the association is reversed for both black and white victims. For,

$$P(Y|A_W V_W) = 19/151 = 0.126 < P(Y|A_B V_W) = 11/63 = 0.175$$

Also,

$$P(Y|A_W V_B) = 0 < P(Y|A_B V_B) = 6/103 = 0.058.$$

Thus, Simpson's paradox occurs.

## 2.2 Simpson's paradox for events

Blyth (1972) first gave the probabilistic interpretation of Simpson's paradox, in terms of conditional probabilities. It may happen that for three events  $A$ ,  $B$  and  $C$ ,

$$P(A|B) < P(A|B^c), \tag{4}$$

while

$$P(A|BC) > P(A|B^cC), \quad P(A|BC^c) > P(A|B^cC^c). \quad (5)$$

As the inequalities in (5) are reversed, compared to (4), the Simpson's paradox occurs.

Since  $P(A|B)$  and  $P(A|B^c)$  are the following weighted averages, namely,

$$\begin{aligned} P(A|B) &= P(A|BC)P(C|B) + P(A|BC^c)P(C^c|B), \\ P(A|B^c) &= P(A|B^cC)P(C|B^c) + P(A|B^cC^c)P(C^c|B^c), \end{aligned} \quad (6)$$

he pointed out that the reversal happens because the weights  $P(C|B)$  and  $P(C^c|B)$  for  $P(A|B)$  are different than the weights  $P(C|B^c)$  and  $P(C^c|B^c)$  for  $P(A|B^c)$ . Note if  $B$  and  $C$  are independent, then the weights for  $P(A|B)$  and  $P(A|B^c)$  are equal and the inequalities of (5) will carry over to  $P(A|B)$  and  $P(A|B^c)$  also. In other words, the Simpson's paradox can not happen in this case. Thus, Simpson's paradox occurs because of the association between  $B$  and  $C$ .

**Remark 1** Look at Example 2 again. The marginal table for the race of the victims (V) and the race of the accused (A) is:

		V	
		W	B
A	W	151	9
	B	63	103

From the above table, the conditional probabilities for the events  $V$  and  $A$  are:

$$\begin{aligned} P(V_W|A_W) &= 0.94; \quad P(V_W|A_B) = 0.38 \\ P(V_B|A_W) &= 0.06; \quad P(V_B|A_B) = 0.62 \end{aligned}$$

showing that there is a strong (marginal) association between  $V$  and  $A$  and leading to Simpson's paradox.

**Remark 2** It is well known that the genesis of Simpson's paradox lies in a simple fact about proportions. There exist positive integers such that  $\frac{k}{l} < \frac{K}{L}$  and  $\frac{m}{n} < \frac{M}{N}$ , but  $\frac{k+m}{l+n} > \frac{K+M}{L+N}$ . For example,  $\frac{1}{6} < \frac{2}{9}$  and  $\frac{5}{7} < \frac{3}{4}$ , but  $\frac{6}{13} > \frac{5}{13}$ . This explains why Simpson's paradox occurs in the analysis of some contingency tables.

## 2.3 Marginal versus conditional association

For the analysis of contingency tables, Lindley and Novick (1981) argue that there is no statistical criterion that would guard against drawing wrong conclusions or would indicate which table (conditional or marginal) represents the correct answer. However, they suggested that if  $C$  is influenced

by  $B$ , then  $C$  should not be treated as a confounding variable. Pearl (1995) also suggested that if  $C$  is affected by  $B$ , then marginal table, rather than the conditional ones, should be used for inference. Thus, causal considerations must be used along with inference. However, there are other researchers who argue Simpson's paradox should not be viewed in terms of causality, as the reversal is real and is not causal. Hence, the paradox is a statistical phenomenon that can be analyzed and avoided using tools of statistical techniques.

One way to avoid Simpson's would be to use a randomized experiment which is not always feasible. Cornfield *et al.* (1959) proposed the minimum effect size criterion to explain an observed association measure  $\rho(A, B)$  between  $A$  and  $B$ , if it is spurious. If  $B$  has no effect or less effect than that of  $C$  on the likelihood of  $A$ , then we would expect

$$\frac{P(C|B)}{P(C|B^c)} > \frac{P(A|B)}{P(A|B^c)},$$

or the risk difference condition, namely,

$$P(A|C) - P(A|C^c) \geq P(A|B) - P(A|B^c).$$

Schild (1999) suggested that this condition could be used as a simple method for deciding whether  $C$  has the strength-the effect size necessary-to reverse the association  $\rho(A, B)$ .

## 2.4 Simpson's paradox as an association reversal phenomena

Samuels (1992) showed that Simpson's paradox between events can be viewed as a particular case of association reversal phenomena for random variables/distributions, which we describe now. Let  $(Y, X, W) \sim F$  and for example  $F_X(x)$  denote the marginal distribution of  $X$ . We say  $W$  is not doubly linked to  $(Y, X)$  if at least one of the following condition holds:

$$(a) W \perp Y, (b) W \perp X, (c) W \perp Y|X, (d) W \perp X|Y.$$

Otherwise, it is doubly linked to  $(Y, X)$ , *i.e.*,  $W$  is linked to both  $Y$  and  $X$ . Henceforth,  $X \perp Y|W$  denotes the conditional independence of  $X$  and  $Y$ , given  $W$ .

An association reversal can be defined for any relation  $R = R(Y, X)$  which denotes the directional association between any two random variables  $X$  and  $Y$ . Henceforth,  $\uparrow$  and  $\downarrow$  means respectively nondecreasing and nonincreasing. Some relations studied in the literature are:

$\mathcal{R}_1$ : (stochastically increasing )  $Y \uparrow X$  if  $P(Y > y|X = x)$  is  $\uparrow$  in  $x$  for all  $y$  .

$\mathcal{R}_2$ : (mean incresing)  $Y \uparrow X$  if  $E(Y|X = x)$  is  $\uparrow$  in  $x$  .

$\mathcal{R}_3$ : (positive quadratic dependence)  $Y \uparrow X$  if  $F(y, x) \geq F_Y(y)F_X(x)$  for all  $(x, y)$

$\mathcal{R}_4$ : (covariance increasing)  $Y \uparrow X$  if  $Cov(X, Y) > 0$  .

The relations  $\mathcal{R}_3$  and  $\mathcal{R}_4$  are symmetric in  $Y$  and  $X$ . The above relations can also be defined for  $\downarrow$  case also.

Samuels (1992) proved that the joint distribution  $F$  cannot exhibit association reversal with respect to  $\mathcal{R}_3$  if  $W$  is not doubly linked to  $(Y, X)$ , that is, when one of the conditions (a) to (d) is true. But, the above result is not true for the relation  $\mathcal{R}_4$ . For example, the condition  $W \perp X|Y$  is not sufficient to prevent the association reversal for  $\mathcal{R}_4$ . However,  $W \perp Y$  prevents association reversal for several  $\mathcal{R}$ 's. See Samuels (1992) for some additional results in this direction.

## 2.5 Linear regression models

Let

$$E(Y|X, W) = \beta_0 + \beta_1 X + \beta_2 W \quad (7)$$

with  $\beta_1 \leq 0$ . Also, let  $\eta = \beta_2 \text{Cov}(X, W)$ . Samuels (1992) showed that the distribution  $F$  exhibits positive association reversal for  $\mathcal{R}_4$  iff  $\eta > 0$ , and  $|\eta| > |\beta_1| \text{Var}(Y)$ . As a corollary, the association reversal with respect to  $\mathcal{R}_2$  holds.

Suppose the marginal model is also linear defined by

$$E(Y|X) = \tilde{\beta}_0 + \tilde{\beta}_1 X,$$

where

$$\tilde{\beta}_1 = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}.$$

Note from (7),

$$\beta_1 = \frac{\text{Cov}(Y, X|W)}{\text{Var}(X|W)}.$$

It is possible that  $\beta_1 < 0$  while  $\tilde{\beta}_1 > 0$ , implying the occurrence of Simpson's paradox for the regression coefficients. Some sufficient conditions for the Simpson's paradox have been recently discussed in Chen, Bengtsson and Ho (2009).

## 2.6 Simpson's paradox in survival analysis

In the context of the survival analysis, it is possible that increasing the value of a covariate  $X$  has a positive effect on a failure time  $T$ , but this effect may be reversed when conditioning on another possible covariate  $Y$ . When studying causal effects and influence of covariates on a failure time  $T$ , this aspect appears paradoxical and creates suspicion on the real effect of  $X$ . These situations may be seen as a kind of Simpson's paradox.

Let  $X$  and  $Y$  be the covariates having effect on  $T$ . Simpson's paradox occurs in survival probability at  $(t, s)$  if

$$\begin{aligned} (a) \quad & P(T > t + s | T > t, X = x, Y = y) \downarrow \text{ in } x, \text{ for all } y \text{ and} \\ (b) \quad & P(T > t + s | T > t, X = x) \uparrow \text{ in } x. \end{aligned} \tag{8}$$

Scarsini and Spizzichino (1999) discussed the Simpson's paradox for different notions of positive dependence and aging.

Let  $h(t|x)$  denote the conditional hazard rate, given  $X = x$ , defined by

$$h(t|x) = \lim_{s \downarrow 0} P(t < T < t + s | T > t; X = x).$$

Then, Simpson's paradox for the hazard rate occurs if

$$h(t|x, y) \uparrow \text{ in } x, \text{ for all } y, \text{ but } h(t|x) \downarrow \text{ in } x. \tag{9}$$

Di Serio, Rinott and Scarsini (2009) showed that Simpson's paradox occurs naturally in the context of survival analysis. They studied the range  $(t, s)$  for which (8) holds, and showed that under certain conditions it holds for all  $t, s > 0$ . They discussed Simpson's paradox for the linear transformation model defined by

$$K(T) = -\beta_x X - \beta_y Y + W, \tag{10}$$

where  $K$  is increasing and  $W \perp (X, Y)$ .

Suppose the covariates  $X$  and  $Y$  satisfy the model

$$Y = \eta(X) + V,$$

where  $\eta$  is  $\uparrow$  and  $X \perp V$ . That is,  $(Y|X = x)$  has density

$$f(y|x) = f_v(y - \eta(x)). \tag{11}$$

Their main result is the following:

**Theorem 1** *Let  $T$  follow the model (10) and the conditional distribution  $\mathcal{L}(Y|X)$  follow (11). Assume  $V$  and  $W$  have increasing failure rates (IFRs). Then,*

*(i) Simpson's paradox in survival probability defined in (8) occurs for all  $t, s > 0$  if*

$$\beta_y < 0 < \beta_x \quad \text{and} \quad \beta_x x + \beta_y \eta(x) \text{ is decreasing in } x. \tag{12}$$

*(ii) If  $W$  and  $V$  are both strictly IFRs, then Simpson's paradox for hazard rate in (9) occurs for all  $t, s > 0$  if and only if (12) holds.*



For example, when  $(Y|X = x) \sim N(\mu + \rho x, 1 - \rho^2)$ , then (9) holds  $\iff \beta_y < 0 < \beta_x$  and  $\rho > \frac{\beta_x}{|\beta_y|}$ .

Note that the linear transformation model corresponds to the survival function

$$\overline{F}_T(t|x, y) = \overline{F}_W(K(t) + \beta_x x + \beta_y y).$$

It can be seen that the Cox's (1972) *proportional hazard model*

$$h(t|x, y) = h_0(t)e^{(\beta_x x + \beta_y y)},$$

where  $h_0(t) > 0$ , is a special case of linear transformation model with  $\overline{F}_W(t) = -e^t$ ,  $t \in \mathbb{R}$ . Also, the *proportional odds model* (Pettitt (1984)) corresponds to the case  $\overline{F}_W(t) = 1/(1 + e^t)$ ,  $t \in \mathbb{R}$ , the *logistic* distribution.

It is interesting to note that even when covariates  $X$  and  $Y$  are independent, there exists the choice of parameters in Cox model for which Simpson's paradox in survival probability occurs for some  $t, s > 0$ . See Di Serio *et al.* (2009) for more details. However, the association or the dependence between  $X$  and  $Y$ , modeled through conditional distributions, is the main source of Simpson's paradox.

## 2.7 Simpson's paradox for an association/dependence measure

Kendall's  $\tau$  and Spearman's  $\rho$  are well known measures of concordance, a certain form of dependence. For example, the dependence of  $Y$  on  $X$  is called stochastically increasing if  $P(Y > y | X = x)$  is increasing in  $x$  for all  $y$ . In other words, when  $X$  is continuous and the partial derivative exists (Cox and Wermuth (2003)), the conditional distribution function  $F(y|x)$  satisfies

$$\frac{\partial F(y | x)}{\partial x} \leq 0, \tag{13}$$

for all  $y$  and  $x$ , with strict inequality in a region of positive probability.

Let  $W$  be a covariate. Then,

$$\frac{\partial F(y | x)}{\partial x} = \int \frac{\partial F(y | x, w)}{\partial x} f(w | x) dw + \int F(y | x, w) \frac{\partial f(w | x)}{\partial x} dw. \tag{14}$$

If  $W \perp X$ , then  $\frac{\partial f(w | x)}{\partial x} = 0$  and hence (Cox (2003))

$$\frac{\partial F(y | x)}{\partial x} = \int \frac{\partial F(y | x, w)}{\partial x} f(w) dw,$$

showing that

$$\frac{\partial F(y | x, w)}{\partial x} \leq 0 \implies \frac{\partial F(y | x)}{\partial x} \leq 0, \text{ for all } y, x \text{ and } w.$$

Thus,  $Y$  remains stochastically increasing in  $x$  after marginalization over the covariate  $W$ . Note in general (see (14)) it is possible that

$$\frac{\partial F(y | x, w)}{\partial x} \leq 0, \text{ for all } (y, x, w), \text{ but } \frac{\partial F(y | x)}{\partial x} > 0$$

for some  $y$  and  $x$ , implying Simpson's paradox.

### 3 Collapsibility

Collapsibility is a concept closely related to that of Simpson's paradox. Generally, whenever Simpson's paradox does not occur, the collapsibility issue arises naturally as a dimension reduction problem in the context of data analysis. It was originally associated with the analysis of contingency tables and so we start with the same.

#### 3.1 Collapsibility of contingency tables

Let  $X_1, \dots, X_n$  be a set of  $n$  categorical variables, where  $X_j \in \{1, \dots, m_j\}$ ,  $1 \leq j \leq n$ . Let  $i = (i_1, \dots, i_n)$  denote a cell of the  $n$ -dimensional table, and  $p(i)$  denote the cell probability with  $p(i) > 0$  and  $\sum_i p(i) = 1$ .

Let  $\bar{n} = \{1, 2, \dots, n\}$ . Define  $l^{(n)}(i) = l^{(n)}(i_1, \dots, i_n) = \ln p(i_1, \dots, i_n)$ . Let  $A = (a_1, \dots, a_r)$ ,  $a_j \in \bar{n}$ ,  $i_A = (i_{a_1}, \dots, i_{a_r})$ , and  $|A|$  denotes the cardinality of  $A$ . Define, for any subset  $A \subset \bar{n}$ ,

$$l_A^{(n)}(i_A) = \sum_{i_j: j \in A^c} l(i_1, \dots, i_n); \quad \tilde{l}_A^{(n)}(i_A) = \frac{1}{\prod_{j \in A^c} m_j} l_A(i_A).$$

For the  $n$ -dimensional table, let

$$l^{(n)}(i) = \sum_{Z \subseteq \bar{n}} \tau_Z^{(n)}(i_Z) \tag{15}$$

be the log-linear model (LLM), where  $\tau_Z^{(n)}(i_Z)$  is the  $r$ -factor interaction parameter when  $|Z| = r$ . Then, it can be seen that (Vellaisamy and Vijay (2007))

$$\tilde{l}_A^{(n)}(i_A) = \sum_{Z \subseteq A} \tau_Z^{(n)}(i_Z), \quad \forall A \subseteq \bar{n}. \tag{16}$$

For example, when  $n=3$ ,

$$\begin{aligned} l^{(3)}(i_1, i_2, i_3) &= \tau_{123}^{(3)}(i_1, i_2, i_3) + \tau_{12}^{(3)}(i_1, i_2) + \tau_{13}^{(3)}(i_1, i_3) + \tau_{23}^{(3)}(i_2, i_3) \\ &\quad + \tau_1^{(3)}(i_1) + \tau_2^{(3)}(i_2) + \tau_3^{(3)}(i_3) + \tau_\phi^{(3)} \\ &= \sum_A \tau_A^{(3)}(i_A), \end{aligned}$$

where  $A$  is any subset of  $\{1, 2, 3\}$ . Then

$$\begin{aligned}\tilde{l}_{12}^{(3)}(i_1, i_2) &:= \frac{1}{m_3} \sum_{i_3} l^{(3)}(i_1, i_2, i_3), \\ &= \tau_{12}^{(3)}(i_1, i_2) + \tau_1^{(3)}(i_1) + \tau_2^{(3)}(i_2) + \tau_\phi^{(3)} \\ &= \sum_{Z \subseteq \{1, 2\}} \tau_Z^{(3)}(i_Z).\end{aligned}$$

Indeed, the interaction factor admits the following representation

$$\tau_A^{(n)}(i_A) = \sum_{Z \subseteq A} (-1)^{|A-Z|} \tilde{l}_Z^{(n)}(i_Z), \quad \forall A \subseteq \bar{n}. \quad (17)$$

Whittemore (1978) defined first  $\tau_A^{(n)}(i_A)$  as a straightforward extension and remarked later that  $l(i) = \sum_{Z \subseteq \bar{n}} \tau_Z(i_Z)$ . Recently, Vellaisamy and Vijay (2007) gave a direct proof of (17) by considering  $\tilde{l}_A$  as the function on the poset  $(\mathcal{P}, \subseteq)$ , and using Möbius inversion theorem. Note also that, by Möbius inversion theorem, (16) holds iff (17) holds.

Let now for simplicity  $A = \{1, \dots, r\}$ , and  $B = \{1, \dots, r, r+1, \dots, s\}$ , where  $r \leq s < n$ . Define  $p_B(i_B) = \sum_{i_j: j \in B^c} p(i)$ , the cell probabilities of the marginal (condensed over  $B^c$ ) table. Define, similarly,

$$l^{(s)}(i) = \ln(p_B(i_B)) = \sum_{Z \subseteq B} \eta_Z^{(s)}(i_Z) \quad (18)$$

be the LLM for the marginal table. Then as seen in the LLM for the full table,

$$\tilde{l}_Z^{(s)}(i_Z) = \frac{1}{\prod_{j \in B \setminus Z} m_j} \sum_{i_j: j \in B \setminus Z} l^{(s)}(i) = \sum_{A \subseteq Z} \eta_A^{(s)}(i_A), \quad (19)$$

for any  $Z \subseteq B$ . The following definition of collapsibility is due to Whittemore (1978).

**Definition 1** An  $n$ -dimensional table is said to be collapsible into an  $s$ -dimensional table with respect to  $\tau_A^{(n)}$ ,  $A \subseteq B$ , if (i)  $\tau_A^{(n)} = \eta_A^{(s)}$  and strictly collapsible if, in addition to (i), (ii)  $\tau_Z^{(n)} = 0$ ,  $\forall Z \supseteq A, Z \cap B^c \neq \emptyset$  holds.

Let

$$d^{(B)}(i_B) = l^{(s)}(i) - \tilde{l}_B^{(n)}(i_B) \quad (20)$$

and for any  $Z \subseteq B$

$$\tilde{d}_Z^{(B)}(i_Z) = \frac{1}{\prod_{j \in B \setminus Z} m_j} \sum_{i_j: j \in B \setminus Z} d^{(B)}(i_B). \quad (21)$$

The next result (Vellaisamy and Vijay (2007)) characterizes the conditions for collapsibility.

**Theorem 2** Let  $\delta_Z = (\eta_Z^{(s)} - \tau_Z^{(n)})$ , for  $Z \subseteq B$ . An  $n$ -dimensional table is collapsible to an  $s$ -dimensional table with respect to  $\tau_A^n$  if and only if

$$\tilde{d}_A^{(B)}(i_A) = \sum_{Z \subseteq A} \delta_Z(i_Z) \iff \sum_{Z \subseteq A} (-1)^{|A-Z|} \tilde{d}_Z^{(B)}(i_Z) = 0, \quad (22)$$

where  $\tilde{d}_Z^{(B)}$  is defined in (21), and  $A \subseteq B$ .

We next mention an important result for the strict collapsibility for hierarchical log-linear models (HLLM), a subclass of LLMs, defined as follows:

**Definition 2** A LLM  $l^{(n)}(i) = \sum_{Z \subseteq \bar{n}} \tau_Z^{(n)}$  is said to be hierarchical if  $\tau_B^{(n)} \neq 0 \implies \tau_A^{(n)} \neq 0$  for  $A \subset B$  or equivalently  $\tau_C^{(n)} = 0 \implies \tau_D^{(n)} = 0$  for  $D \supset C$ .

Let now  $\bar{n} = A + B + C$ . For a HLLM, Bishop, Fienberg and Holland (1975) (BFH (1975)) showed that the  $n$ -dimensional table is collapsible into a  $s$ -dimensional table (over  $C$ ) with respect to  $\tau_{AUV}^{(n)}$ , where  $V \subseteq B$ , iff  $\tau_Z^{(n)} = 0$ , for all  $Z \cap A \neq \phi$  and  $Z \cap B \neq \phi$ , that is,  $X_A \perp X_C | X_B$ . Later, Whittemore (1978) showed that they are only sufficient but not necessary. Recently, Vellaisamy and Vijay (2007) showed that those conditions are necessary and sufficient for strict collapsibility with respect to a set of interaction parameters, which is stated below.

**Theorem 3** Let  $\bar{n} = A + B + C$  be such that  $|A \cup B| = s$  and  $|C| = n - s$ . Then, an  $n$ -dimensional table is strictly collapsible (over  $C$ ) into an  $s$ -dimensional table with respect to the set  $C_L = \{\tau_L | L \subseteq A \cup B; L \cap A \neq \phi\}$  if and only if  $X_A \perp X_C | X_B$ .

It follows from the above result that for  $k \in \{1, 2\}$ , a 3-dimensional table is strictly collapsible into a 2-dimensional table with respect to  $\tau_k^{(3)}$  and  $\tau_{12}^{(3)}$  iff  $\tau_{123}^{(3)} = 0$  and  $\tau_{k3}^{(3)} = 0$ . Note also that when  $k=1$ , the conditions  $\tau_{123}^{(3)} = 0$  and  $\tau_{13}^{(3)} = 0$  are nothing but BHF's (1975) sufficient conditions for collapsibility with respect to  $\tau_{12}^{(3)}$  or  $\tau_{23}^{(3)}$ .

For some recent results on the collapsibility of full tables based on conditional tables/models, one may refer to Vellaisamy and Vijay (2010). The concept of collapsibility for contingency tables was later extended to the study of regression models by Wermuth (1989) and several others.

### 3.2 Collapsibility of regression coefficients

Let  $Y$  be a continuous response variable,  $X$  be a continuous influence variable and  $A$  be a discrete (background) variable with levels  $i = 1, 2, \dots, I$ . Initially, the problems of collapsibility were addressed only for parallel regression models (Wermuth (1989)) defined by

$$E(Y|X = x, A = i) = \alpha_{yx}(i) + \beta_{yx}x, \quad 1 \leq i \leq I, \quad (23)$$

where  $\beta_{yx} = \frac{\sigma_{yx}(i)}{\sigma_{xx}(i)} = \frac{Cov(Y, X|A=i)}{V(X|A=i)}$  is the regression coefficient. Since  $\sigma_{yx}(i) = \beta_{yx}\sigma_{xx}(i)$  for all  $i$ , we have

$$\sigma_{yx}(A) = Cov(Y, X|A) = \beta_{yx}V(X|A).$$

Let us now introduce the following notation:

$$\begin{aligned}\mu_y(A) &= E(Y|A), & \mu_x(A) &= E(X|A), & \sigma_{xx}(A) &= V(X|A), \\ \sigma_{yx} &= Cov(Y, X), & \sigma_{xx} &= V(X), & P(A=i) &= \pi_i > 0.\end{aligned}$$

In general,  $\beta_{yx}(A) = (\sigma_{yx}(A)/\sigma_{xx}(A))$  is a function of  $A$ . In the parallel regression model,  $\beta_{yx}(A) = \beta_{yx}$ . Note also from the model (23),  $\mu_y(A) = E_{X|A}(\alpha_{yx}(A) + \beta_{yx}X)$ . The following definition of collapsibility is due to Wermuth (1989).

**Definition 3** *The parallel regression coefficient  $\beta_{yx}$  is said to be collapsible over  $A$  if  $\beta_{yx} = \tilde{\beta}_{yx}$ , where  $\tilde{\beta}_{yx}$  is the regression coefficient for the marginal linear model*

$$E(Y|X=x) = \tilde{\alpha}_{yx} + \tilde{\beta}_{yx}x. \quad (24)$$

The above model implies  $\sigma_{yx} = \tilde{\beta}_{yx}\sigma_{xx}$ . Next we present a necessary and sufficient condition for collapsibility, due to Wermuth (1989) (see also Vellaisamy and Vijay (2008) for a simple probabilistic proof).

**Theorem 4** *The regression coefficient  $\beta_{yx}$  of the parallel regression model (23) is collapsible over  $A$  if and only if  $Cov_A(\alpha_{yx}(A), \mu_x(A)) = 0$ .*

As a corollary, (i)  $\alpha_{yx}(A)$  or  $\mu_x(A)$  is degenerate, or (ii)  $\beta_{yx} = (\mu_y(A)/\mu_x(A))$  a.e., with  $\mu_x(A) \neq 0$ , is a sufficient condition for collapsibility.

The next result, due to Vellaisamy and Vijay (2008), is more general, as it does not assume  $(Y, X, A)$  follows a conditional Gaussian distribution, a condition usually assumed in the literature.

**Theorem 5** *The regression coefficient  $\beta_{yx}$  of the model (23) is collapsible if*

- (i)  $Y \perp A|X$  or
- (ii)  $X \perp A|Y$  and  $V_A(\mu_y(A))E_A(\sigma_{yy}(A)) = V_A(\mu_x(A))E_A(\sigma_{xx}(A))$ .

**Remark 3** *Note that if  $X \perp A|Y$  and  $Y \perp A$ , then  $A \perp (X, Y)$  (Whittaker (1990)) and hence condition (ii) of the above result is satisfied. Thus,  $\beta_{yx}$  is collapsible.*

### 3.3 Random Coefficient Regression Models

The condition that  $Cov((Y, X)|A = i)/V(X|A = i)$  is independent of the levels of  $A$  is stronger and may not hold in several real-life applications. For example, the well-known degradation models of the form

$$y_i(t) = \alpha(i) - \beta(i)t, \quad (25)$$

where  $y_i(t)$  denotes the log-performance of specimen  $A = i$  as a function of age  $t$ , shows that different specimens have different linear degradation. Such models arise in accelerated life-testing problems (Nelson (2004), p. 530). The model (25) can be written in the form

$$E(Y|X = x, A = i) = \alpha_{yx}(i) + \beta_{yx}(i)x,$$

or equivalently

$$E(Y|X, A) = \alpha_{yx}(A) + \beta_{yx}(A)X, \quad (26)$$

where  $\beta_{yx}(A) = (\sigma_{yx}(A)/\sigma_{xx}(A))$ , is a random coefficient regression model. For these models, Vellaisamy and Vijay (2008) introduced and studied average collapsibility which we discuss next.

**Definition 4** *The random regression coefficient  $\beta_{yx}(A)$  is said to be average collapsible ( $A$ -collapsible) if  $\tilde{\beta}_{yx} = E_A(\beta_{yx}(A))$ , where  $\tilde{\beta}_{yx}$  is the regression coefficient of the marginal linear model*

$$E(Y|X = x) = \tilde{\alpha}_{yx} + \tilde{\beta}_{yx}x. \quad (27)$$

The following result generalizes Theorem 4.

**Theorem 6** *The random regression coefficient  $\beta_{yx}(A)$  of the model (26) is  $A$ -collapsible if and only if*

$$E_A(\beta_{yx}(A))V(\mu_x(A)) = Cov(\beta_{yx}(A), \sigma_{xx}(A)) + Cov(\mu_y(A), \mu_x(A)). \quad (28)$$

It is of practical interest to know the conditions under which both the random regression coefficients are collapsible.

**Theorem 7** *Consider the random coefficients regression model (26) with  $P(X = 0) = 0$ . Then  $\alpha(A)$  and  $\beta(A)$  are both  $A$ -collapsible if one of the following conditions holds:*

$$(i) \quad E(\alpha(A)|X) = \tilde{\alpha} \text{ a.e.}, \quad (ii) \quad E(Y|X, A) = E(Y|X) \text{ a.e.}$$

Next, we briefly discuss the collapsibility problems for the logistic regression coefficients. Let  $Y$  be a binary response variable taking the values 0 and 1,  $X$  be a random vector of  $p$  risk factors and  $A$  be a discrete background variable with levels  $i = 1, \dots, I$ . Guo and Geng (1995) obtained some collapsibility results for the regression coefficients of the logistic regression model. We focus only on random coefficient logistic regression model of  $Y$  on  $X$ , for the levels of  $A$ , defined by

$$\ln\left\{\frac{P(Y=1|X, A)}{P(Y=0|X, A)}\right\} = \alpha(A) + \beta^T(A)X. \quad (29)$$

We say that the logistic regression coefficient vectors  $\beta(A)$  is  $A$ -collapsible if  $E_A(\beta(A)) = \tilde{\beta}$ , where  $\tilde{\beta}$  is the regression coefficient vector for the marginal regression model

$$\ln\left\{\frac{P(Y=1|X)}{P(Y=0|X)}\right\} = \tilde{\alpha} + \tilde{\beta}^T X.$$

**Theorem 8** *Let  $X$  be a continuous random vector. Then, for the model (29),*

- (i)  $A \perp Y|X$  implies  $\alpha(A)$  and  $\beta(A)$  both are  $A$ -collapsible.
- (ii)  $A \perp X|Y$  implies  $\beta(A)$  is  $A$ -collapsible.

For a proof, see Vellaisamy and Vijay (2008). Finally, we address the collapsibility issues for a certain measure of dependence for two random variables.

### 3.4 Collapsibility of distribution dependence

Let  $F(y|x, w)$  denote the conditional distribution function, where  $Y$  is a response variable,  $X$  is an explanatory variable (continuous) and  $W$  is a background variable. Then, the function  $\frac{\partial F(y|x, w)}{\partial x}$ , when it exists, is called a distribution dependence function (Cox and Wermuth (2003)). It represents the stochastically increasing property between  $X$  and  $Y$ . When  $X$  is discrete, the partial differentiation is replaced by differencing between adjacent levels of  $X$ . The following definition is due to Ma, Xie and Geng (2006).

**Definition 5** *The distribution dependence function is said to be homogeneous with respect to  $W$  if  $\frac{\partial F(y|x, w)}{\partial x} = \frac{\partial F(y|x, w')}{\partial x}$ , for all  $y, x$  and  $w \neq w'$  and collapsible over  $W$  if  $\frac{\partial F(y|x, w)}{\partial x} = \frac{\partial F(y|x)}{\partial x}$ , for all  $y, x$  and  $w$ .*

Ma *et al.* (2006) showed that the distribution dependence function is uniformly collapsible and hence collapsible iff either (a)  $Y \perp X|W$ ; or (b)  $X \perp W$  and  $\frac{\partial F(y|x, w)}{\partial x}$  is homogeneous in  $w$ . Cox and Wermuth (2003) showed that either condition (a) or (b) is sufficient to ensure that no effect

reversal or Simpson's paradox occurs. Note that homogeneity is a stronger condition which may not hold for most of the models that are encountered in practice. For example, consider a simple linear regression model defined by  $Y = m(X, W) + \epsilon$ , where  $m(x, w) = \alpha_1 x + \alpha_3 xw$ , and  $\epsilon \sim N(0, 1)$ . Let  $\phi$  be the standard normal density. Then,

$$\frac{\partial F(y|x, w)}{\partial x} = -(\alpha_1 + \alpha_3 w)\phi(y - m(x, w)),$$

and hence is not homogeneous over  $W$ . For such models, the concept of average collapsibility introduced and studied by Vellaisamy (2011) is a very useful concept. Indeed, when the distribution dependence function is homogeneous, it reduces to collapsibility. We say that the distribution dependence function  $\frac{\partial F(y|x, w)}{\partial x}$  is average collapsible over  $W$  if

$$E_{W|X=x} \left( \frac{\partial F(y|x, W)}{\partial x} \right) = \frac{\partial F(y|x)}{\partial x}, \text{ for all } y \text{ and } x.$$

Vellaisamy (2011) showed that average collapsibility holds if (i)  $Y \perp W | X$  or (ii)  $W \perp X$  holds. These conditions are also necessary when  $W$  is a binary variable. An example, where average collapsibility holds, follows next. Let  $\phi(z)$  and  $\Phi(z)$  respectively denote the density and the distribution of  $Z \sim N(0, 1)$ .

**Example 3** Consider the linear regression model

$$Y = \alpha_1 X + \alpha_2 W + \alpha_3 XW + \epsilon,$$

where  $\epsilon \perp (X, W)$  and  $\epsilon \sim N(0, \sigma^2)$ . Then

$$(Y|x, w) \sim N(m(x, w), \sigma^2),$$

where  $m(x, w) = \alpha_1 x + \alpha_2 w + \alpha_3 xw$ . Hence,

$$\frac{\partial F(y|x, w)}{\partial x} = \left(\frac{-1}{\sigma}\right)(\alpha_1 + \alpha_3 w)\phi\left(\frac{y - m(x, w)}{\sigma}\right),$$

which is not homogeneous.

Suppose  $W \sim N(0, 1)$  and  $W \perp X$ . Then  $(Y|x) \sim N(\alpha_1 x, v^2(x, \sigma))$ , where  $v^2(x, \sigma) = (\alpha_2 + \alpha_3 x)^2 + \sigma^2$ . Then, it can be shown that (see Vellaisamy (2011))

$$E_{W|X=x} \left( \frac{\partial F(y|x, W)}{\partial x} \right) = \frac{\partial F(y|x)}{\partial x},$$

so that average collapsibility over  $W$  holds.



Note also from (14) that average collapsibility holds if and only if

$$\int F(y | x, w) \frac{\partial f(w | x)}{\partial x} dw = 0 \quad \text{for all } (y, x). \quad (30)$$

The conditions (i) and (ii) of average collapsibility are not necessary, unless  $W$  is binary. A counter-example follows:

**Example 4** Let  $(Y|x, w) \sim U(0, (x^2 + (w - x)^2)^{-1})$  so that

$$F(y|x, w) = y(x^2 + (w - x)^2), \quad 0 < y < (x^2 + (w - x)^2)^{-1}.$$

Assume also  $(W|X = x) \sim N(x, 1)$  so that

$$\frac{\partial}{\partial x} f(w|x) = -\phi'(w - x) = (w - x)\phi(w - x).$$

Then it can be seen that (see Vellaisamy (2011))

$$\int F(y|x, w) \frac{\partial}{\partial x} f(w|x) dw = 0, \quad \text{for all } (y, x).$$

Thus, from (30), average collapsibility over  $W$  holds, but neither condition (i) nor condition (ii) is satisfied.

**Conclusions.** The examples and the applications discussed in this paper clearly demonstrate that Simpson's paradox is a crucial aspect in the data analysis and the issue of collapsibility should be looked into only after ascertaining the nonoccurrence of Simpson's paradox. Only recently, the issue of Simpson's paradox for survival analysis and for certain measures of association has been addressed. However, the conditions of collapsibility for survival models, when the co-variate is either known or unknown, are yet to be explored. Specifically, the concept of average collapsibility (Vellaisamy (2011)) is more relevant in view of the nature of the Simpson's paradox in survival models (Di Serio *et. al* (2009)). These and other considerations are of practical interest and some of these issues are already under consideration. The findings will be reported elsewhere.

**Acknowledgements.** This work was completed while the author was visiting the Department of Statistics and Probability, Michigan State University, USA. The author is grateful to Professor Hira L. Koul for all the support and encouragement for the timely completion of this work, and for some helpful comments which improved the presentation of the paper. This research is partially supported by a DST research grant No. SR/S4/MS: 706/10.

## References

- Bickel, P. J., Hjammel, E. A. and O' Connel, J. W. (1975). Sex bias in Graduate admissions: Data from Berkeley. *Science*, **187**, 398-404.
- Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.
- Blyth, C.R. (1973). Simpson's paradox and mutually favourable events. *J. Amer. Statist. Assoc.*, **68**, 746.
- Chen, A., Bengtsson, T. and Ho, T.K. (2009). A regression paradox for linear models: sufficient conditions and relation to Simpson's paradox. *Amer. Statistician*, **63**, 218-225.
- Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin M., *et al.* (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. *J. Nat. Cancer Institute*, **22**, 173-203.
- Cox, D. R. (2003). Conditional and marginal association for binary random variables. *Biometrika*, **90**, 982-984.
- Cox, D.R. and Wermuth, N. (2003). A general condition for avoiding effect reversal after marginalization. *J. R. Statist. Soc. B*, **65**, 937-941.
- Cox, D. R. (1972). Regression models and life-tables. *J. R. Statist. Soc. B*, **34**, 187-220.
- Di Serio, C., Rinott, Y. and Scarsini, M. (2009). Simpson's paradox in survival models . *Scand. J. Statist.*, **36**, 463-480.
- Geng, Z. and Asano, C. (1993). Strong collapsibility of association measures in linear models. *J. R. Statist. Soc. B*, **55**, 741-747.
- Guo, J. H. and Geng, Z. (1995). Collapsibility of logistic regression coefficients. *J. R. Statist. Soc. B*, **57**, 263-267.
- Lindley, D.V. and Novick, M.R. (1981). On the role of exchangeability in inference. *Ann. Statist.*, **9**, 45-58.
- Ma, Z., Xie, X. and Geng, Z. (2006). Collapsibility of distribution dependence. *J. R. Statist. Soc. B*, **68**, 127-133.
- Nelson, W. B. (2004). *Accelerated Testing: Statistical Models, Test Plans and Data Analysis*. John Wiley and Sons, New Jersey.

- Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 110-125.
- Pettitt, A. N. (1984). Proportional odds model for survival data and estimates using ranks. *Appl. Statist.*, **33**, 169-175.
- Samuels, M.L. (1993). Simpson's paradox and related phenomena. *J. Amer. Statist. Assoc.*, **88**, 81-88.
- Scarsini, M. and Spizzichino, F. (1999). Simpson-type paradoxes, dependence, and ageing. *J. Appl. Probab.*, **36**, 119-131.
- Simpson, E.H. (1951). The interpretation of interaction in contingency tables. *J. R. Statist. Soc. B*, **13**, 238-241.
- Schild, M. (1999). Simpson's paradox and Cornfield's conditions. *Proceedings of the ASA-JSM Section of Statistical Education*, ASA, 106-111.
- Yule, G.U. (1903). Notes on the theory of association of attributes. *Biometrika*, **2**, 121-134.
- Vellaisamy, P. and Vijay, V. (2007). Some collapsibility results for n-dimensional contingency tables. *Ann. Inst. Statist. Math.*, **59**, 557-576.
- Vellaisamy, P. and Vijay, V. (2008). Collapsibility of regression coefficients and its extensions. *J. Statist. Plann. Inference*, **138**, 982-994.
- Vellaisamy, P. and Vijay, V. (2010). Collapsibility of contingency tables based on conditional models. *J. Statist. Plann. Inference*, **140**, 1243-1255.
- Vellaisamy, P. (2011). Average collapsibility of distribution dependence and quantile regression coefficients. To appear in *Scand. J. Statist.*
- Wermuth, N. (1987). Parametric collapsibility and the lack of moderating effects in contingency tables with a dichotomous response variable. *J. R. Statist. Soc. B*, **49**, 353-364.
- Wermuth, N. (1989). Moderating effects of subgroups in linear models. *Biometrika*, **76**, 81-92.
- Whittemore, A. S. (1978). Collapsibility of multidimensional contingency tables. *J. R. Statist. Soc. B*, **40**, 328-340.
- Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*, John Wiley and Sons, New York.
- Yule, G.U. (1903). Notes on the theory of association of attributes. *Biometrika*, **2**, 121-134.